AMENDMENTS TO THE CLAIMS

1-16 (Cancelled)

17. A method for modifying a collection of inverted term lists, comprising:

(a) creating a compressed document surrogate for each document in a database which the collection of inverted term lists summarizes, the compressed document surrogate for a document containing information to identify each term which occurs in the document for which there is an inverted term list; and,

(b) updating the collection of inverted term lists, when a document in the database which the collection of inverted term lists summarizes is modified or deleted, by using the compressed document surrogate for the document to determine terms for which there are inverted term lists in the document, and updating the inverted term lists to reflect: a modification or deletion, and terms added to the document that were not previously in the document.

18. The method of claim 16, where the information about each term included in a compressed document surrogate for a document, includes at least one of: a term identification number, a location in a lookup table of an entry for the term, an address of an inverted term list of the term, an address of a location in the inverted term list for the term of the document, a number of times the term occurs in the document, and a location in the document of each occurrence of the term.

19. The method of claim 17, where information about the terms is stored in the compressed document surrogate in term identification number order, and the term identification number of a term in the compressed document surrogate is given relative to the term identification number of a prior term in the document.

20. A method for maintaining a database with information about a collection of documents, to facilitate determining which documents may be of interest, where the documents in the collection may be modified or deleted from time to time, the method comprising:

(a) choosing some or all of the terms found in the collection of documents to be indexed;

(b) for each term chosen, preparing an inverted term list or inverted term lists, said list or lists containing information about the term's occurrence in the collection of documents;

(c) for each document in the collection, preparing a compressed document surrogate, said surrogate comprising a list of each term, for which there is an inverted term list, which occurs in the document, together with additional information about the occurrence of said term in said document;

(d) when a document in the collection which the set of inverted term lists summarizes is modified or deleted, updating the set of inverted term lists, by consulting the compressed document surrogate for said document to determine which terms for which there are inverted term lists in said document, updating the inverted term lists corresponding to those terms to reflect the modification or deletion in the document, and updating the inverted term lists to reflect terms added to the document that were not previously in the document; and,

(e) specifying at least one of desired terms which are to be found in documents and undesired terms which are not to be found in documents, determining those documents containing at least one of the desired terms and the undesired terms, and how often the desired terms or the undesired terms appear in the documents, by consulting the inverted term lists for the desired and the undesired terms, and preparing a list of documents ordered depending upon the occurrence of at least one of the desired terms and the undesired terms.

21. The method of claim 19, wherein a unique term identification number is assigned to each term, and a compressed document surrogate contains the term identification number of each term contained in the document.

22. The method of claim 20, wherein the information is stored in the compressed document surrogate in order of term identification number, and the term identification number of a term is given relative to the prior term identification number.

23. The method of claim 20, wherein a compressed document surrogate contains the number of times the term occurs in the document.

24. The method of claim 19, wherein the inverted term lists contain a document identification number for each document in which the term appears, and the number of times the term occurs in the document.

25. The method of claim 23, wherein documents are listed in an inverted term list in order of their term frequency scores.

26. The method of claim 19, wherein two inverted term lists are maintained for each term, a top inverted term list containing information about the documents in which the term occurs most frequently, and a remainder inverted term list containing information about the remaining documents in which the term occurs.

27. The method of claim 19, wherein a lookup table maintains, for each term, the term in a natural language, the address of each inverted term list for the term, the number of documents containing the term, and numbers reflecting the maximum amount the term can contribute to the score of a document on each of the term's inverted term lists, when processing a search query.

28. The method of claim 26, wherein the lookup table is a fixed array with information about terms stored in order of term identification numbers.

29 The method of claim 19, wherein an inverted term list for a term contains information about the location within a document of each occurrence of the term in question.

30 The method of claim 28, wherein the locations within the document of each occurrence of the term in question are given in relation to the prior occurrence of the term in the document.

31 The method of claim 19, wherein the documents are Web pages.

32 The method of claim 19, wherein the documents are Web sites.

33 A method of determining the score for a document under a search query which specifies terms that are to be present or absent, the method comprising:

(a) creating a compressed document surrogate for each document in the database, the compressed document surrogate contains information about each term, from among the terms of interest of interest in the database, which occurs in the document, and which compressed document surrogate is created with inverted term lists that contain information about the terms of interest in the database, and where the information about each term included in the compressed document surrogate for a document includes at least one of: the term identification number of the term, the location in a lookup table of an entry for the term, the number of times the term occurs in the document, the location in the document of each occurrence of the term, the address of the inverted term list of the term, and the address of the location in the inverted term list for the term of the document;

(b) consulting the compressed document surrogate for the document whose score is to be determined;

(c) for each term contained in said compressed document surrogate, consulting at least one of an associated inverted term list and a lookup table, and calculating the contribution to the document score resulting from said term; and,

(d) determining the total document score by adding the contributions of each term in the compressed document surrogate.

34. The method of claim 32, where at (c) the inverted term list is not consulted.

35. The method of claim 32, wherein the documents are Web pages.

36. The method of claim 32, wherein the documents are Web sites.

37. A method for returning a list of a number of documents N in order of predicted utility, from among a collection of documents, as predicted by a search query containing terms to be present or absent, the method comprising:

(a) creating a compressed document surrogate for each document in the database, where the compressed document surrogate contains information about each term, from among the terms of interest of interest in the database, which occurs in the document, and which compressed document surrogate is created with top and remainder inverted term lists that contain information about the terms of interest in the database, and where the information about each term included in the compressed document surrogate for a document includes at least one of: the term identification number of the term, the location in a lookup table of an entry for the term, the number of times the term occurs in the document, the location in the document of each occurrence of the term, the address of the inverted term list of the term which contains the document, and the address of the location in the inverted term list of the document;

(b) choosing, from among the terms in the search query which are to be found in documents, the term whose top inverted term list has not yet considered, which occurs in the fewest documents in the collection;

(c) consulting the top inverted term list for said term, calculating the score for each document found in the top inverted term list;

(i) if the document has not previously been found on an inverted term list, assigning the document the calculated score;

(ii) if the document has previously been found on an inverted term list, increasing its previously-calculated score by the calculated score;

(d) calculating a maximum score, S_{Max}, achieved by a document, not already found on a top inverted term list, if it is found on all top inverted term lists, for terms to be found in documents, not yet consulted;

(e) calculating a maximum score, S_{Sub}, to be subtracted from a document score, as a result of said document being found to contain terms to be absent from a document;

(f) determining whether there are N or more documents already found, with scores such that if S_{Sub} were subtracted from their scores, the remainder would be greater than S_{Max};

(g) if there are N or more such documents, determining by use of the compressed document surrogate for each document a final score for the documents that have so far been found in any inverted term list of a desired term, and providing a list of the N documents with the highest scores, ranked in order of score;

(h) if there are not N or more such documents, repeating (b) through (f) until either N or more such documents are found, or until no top inverted term list of a term to be found in the document has not been analyzed;

(i) if there are not N or more such documents, and the top inverted term lists of all terms desired to be found in the document have been analyzed, repeating (b)

through (h) utilizing remainder inverted term lists instead of top inverted term lists, until either N or more such documents are found, or until no remainder inverted term lists of terms desired to be found in the document has not been analyzed; and,

(j) determining by use of the compressed document surrogate for each document the final score for the documents found on the inverted term lists of the desired terms, and providing a list of the documents ranked in order of score.

38. The method of claim 36, wherein the documents are Web pages.

39. The method of claim 36, wherein the documents are Web sites.

40. The method of claim 36, wherein only terms desired to be found are contained in a search query, so that S_{Sub} is zero.

41. A method for choosing documents of interest from a collection of documents, the method comprising:

(a) determining an initial selection criterion;

(b) applying the initial selection criterion to each document in the collection, to generate a rank-ordered list of documents;

(c) evaluating a subset of the documents on the list to determine relevant documents and irrelevant documents;

(d) modifying the selection criteria by at least one of: adjusting weights assigned to each element of the selection criteria in the prior iteration, removing elements of the selection criteria from the prior iteration, and adding additional elements to the selection criteria, based upon features of the relevant documents, by use of compressed document surrogates for the relevant documents, where said compressed document surrogates comprise information about the use of terms in the relevant documents;

(e) applying the modified selection criterion to each document in the collection, to generate a new rank-ordered list of documents; and,

(f) repeating (c), (d), and (e).

42. The method of claim 40, wherein when the modified selection criterion are applied to each document in the collection at (e), to generate a new rank-ordered list of documents; the compressed document surrogates for the documents are utilized to calculate the final document scores.

43. The method of claim 40, wherein the documents classified are Web pages.

44. The method of claim 40, wherein the documents classified are Web sites.

45. The method of claim 40, wherein the initial selection criteria are arbitrarily chosen.

46. The method of claim 40, wherein the documents classified are one of: electronic commerce Web pages and electronic commerce Web sites.

47. The method of claim 41, wherein modifying the selection criteria at (d) includes at least one of: adjusting a weight assigned to each element of the selection criteria in the prior iteration, removing elements of the selection criteria in the prior iteration, and adding additional elements to the criteria, based upon features of the relevant documents and irrelevant documents, by use of compressed document surrogates for the relevant documents and the irrelevant documents, where said compressed document surrogates comprise information about the use of terms in the relevant documents and the irrelevant documents.

48. The method of claim 47, wherein modifying the selection criteria includes:

(a) giving each term found in the collection of documents a score based upon how often the term occurs in relevant documents, compared to how often the term occurs in the collection of documents as a whole, and based upon how often the term occurs in the irrelevant documents, compared to how often the term occurs in the collection of documents as a whole;

(b) choosing terms with the highest positive weights thus determined to be the terms in the selection criteria; and,

(c) weighing the terms in the selection criteria according to the scores achieved in the above process, and the relative frequency of the terms in the collection.

49. The method of claim 48 wherein a score W_T given to a Term T at (a) is determined by a formula:

$W_T = \log(P_T(R)/P_T(\underline{R}))$, where

$P_T(R)$ = a probability that the term T occurs in a relevant document,

$= N_{TR}/(\Sigma_R N_{tR})$, where:

N_{TR} = a number of occurrences of the term T in a relevant document,

$\Sigma_R N_{tR}$ = a total number of occurrences of terms in relevant documents,

$P_t(\underline{R})$ = a probability that the term T occurs in an irrelevant document,

$= N_{T\underline{R}}/(\Sigma_R N_{t\underline{R}})$, where

$N_{T\underline{R}}$ = a number of occurrences of the term T in irrelevant documents,

$\Sigma_{\underline{R}} N_{t\underline{R}}$ = a total number of occurrences of terms in irrelevant documents.

50. The method of claim 49, wherein the terms chosen at (b) are the terms whose scores W_T exceed an average score W_T by two or more standard deviations.

51. The method of claim 50, wherein weights S_T assigned to terms at (c) are determined by a formula:

$$S_T = W_T * IDF_T,$$

where: $IDF_T = \log((N+K_3)/N_T)/\log(N + K_4)$

where:

N is a number of documents in the subset,

N_T is a number of documents containing the term T in the subset,

K_3 and K_4 are constants.

52. The method of claim 51, wherein K_3 is 0.5, and K_4 is 1.0.

53. The method of claim 51, wherein in applying the modified selection criterion to each document in the collection, to generate a new rank-ordered list of documents, documents are ranked in order of their scores S_D,

where: $S_D = \Sigma(S_T * TF_{TD})$,

S_T has the value set forth above,

TF_{TD} = Robertson's term frequency for Term T in Document D

$= NT_D/(NT_D + K_1 + K_2 * (L_D/L_0))$,

where: NTD is a number of times the term T occurs in document D,

L_D is a length of document D,

L_0 is an average length of a document in the subset of documents indexed, and

K_1 and K_2 are constants.

54. The method of claim 53, wherein K_1 is 0.5, and K_2 is 1.5.

55. A method for identifying documents in a collection as having a particular characteristic, the method comprising:

(a) choosing an initial list of documents from among the documents in the collection;

(b) evaluating a subset of the documents on the list to determine whether each document in the subset has the characteristic;

(c) modifying the selection criteria by at least one of: adjusting the weights assigned to each element of the selection criteria in the prior iteration, removing elements of the selection criteria in the prior iteration, and adding additional elements to the criteria, based upon features of the documents determined to have the characteristic, and based upon features of the documents determined not to have the characteristic, by use of compressed document surrogates for the documents, where said compressed document surrogates comprise information about the use of the terms in the documents found to have the characteristic or not to have the characteristic;

(d) applying the modified selection criterion to each document in the initial list of documents, to generate a new rank-ordered list of documents;

(e) repeating (b), (c), and (d);

(f) choosing a cutoff score to be applied; and,

(g) concluding that all documents in the collection with scores above the cutoff score have the characteristic.

56. The method of claim 55, wherein modifying the selection criteria at (c) includes:

(a) giving each term found in the subset of documents a score based upon how often the term occurs in documents determined to have the characteristic, compared to how often the term occurs in the subset of documents as a whole, and based upon how often the term occurs in documents determined not to have the characteristic, compared to how often the term occurs in the subset of documents as a whole;

(b) choosing terms with the highest positive weights determined to be the terms in the selection criteria; and,

(c) weighing the terms in the selection criteria according to the scores achieved in the above process, and their relative frequency in the subset.

57. The method of claim 56, wherein a score W_T given to a Term T at (a) is determined by a formula:

$W_T = \log(P_T(R)/P_T(\underline{R}))$, where

$P_T(R)$ = a probability that the term T occurs in a document determined to have the characteristic,

$= N_{TR}/(\Sigma_R N_{tR})$, where:

N_{TR} = a number of occurrences of the term T in a document determined to have the characteristic,

$\Sigma_R N_{tR}$ = a total number of occurrences of terms in a document determined to have the characteristic,

$P_t(\underline{R})$ = a probability that the term T occurs in a document determined not to have the characteristic,

$= N_{T\underline{R}}/(\Sigma_R N_{t\underline{R}})$, where

$N_{T\underline{R}}$ = a number of occurrences of the term T in documents determined not to have the characteristic,

$\Sigma_{\underline{R}} N_{t\underline{R}}$ = a total number of occurrences of terms in documents determined not to have the characteristic.

58. The method of claim 57, wherein the terms chosen at (b) are the terms whose scores W_T exceed an average score W_T by two or more standard deviations.

59. The method of claim 58, wherein weights S_T assigned to the terms at (c) are determined by a formula:

$S_T = W_T * IDF_T$,

where: $IDF_T = \log((N+K_3)/N_T)/\log(N + K_4)$

where:

N is a number of documents in the subset,

N_T is a number of documents containing the term T in the subset,

K_3 and K_4 are constants.

60. The method of claim 59, wherein K_3 is 0.5, and K_4 is 1.0.

61. The method of claim 59, wherein in applying the modified selection criterion to each document in the subset, to generate a new rank-ordered list of documents, documents are ranked in order of their scores S_D,

where: $S_D = \Sigma(S_T * TF_{TD})$,

S_T has the value set forth above,

TF_{TD} = Robertson's term frequency for Term T in Document D

$= NT_D/(NT_D + K_1 + K_2*(L_D/L_0))$,

where: NTD is a number of times the term T occurs in document D,

L_D is a length of document D,

L_0 is an average length of a document in the subset of documents indexed, and

K_1 and K_2 are constants.

62. The method of claim 61, wherein K_1 is 0.5, and K_2 is 1.5.

63. The method of claim 61, where the documents are Web pages.

64. The method of claim 61, where the documents are Web sites.

65. The method of claim 64, where the particular characteristic is being an electronic commerce site.

66-81 (Cancelled)

82. A device for modifying a collection of inverted term lists, the device comprising:

(a) means for creating a compressed document surrogate for each document in a database which the collection of inverted term lists summarizes, the compressed document surrogate for a document containing information to identify each term which occurs in the document for which there is an inverted term list; and,

(b) means for updating the collection of inverted term lists, when a document in the database which the collection of inverted term lists summarizes is modified or deleted, by using the compressed document surrogate for the document to determine terms for which there are inverted term lists in the document, and updating the inverted term lists to reflect: a modification or deletion, and terms added to the document that were not previously in the document.

83. The device of claim 82, where the information about each term included in a compressed document surrogate for a document, includes at least one of: a term identification number, a location in a lookup table of an entry for the term, an address of an inverted term list of the term, an address of a location in the inverted term list for the term of the document, a number of times the term occurs in the document, and a location in the document of each occurrence of the term.

84. The device of claim 83, where information about the terms is stored in the compressed document surrogate in term identification number order, and the term identification number of a term in the compressed document surrogate is given relative to the term identification number of a prior term in the document.

85. A device for maintaining a database with information about a collection of documents, to facilitate determining which documents may be of interest, where the documents in the collection may be modified or deleted from time to time, the device comprising:

(a) means for choosing some or all of the terms found in the collection of documents to be indexed;

(b) means for preparing, for each term chosen, an inverted term list or inverted term lists, said list or lists containing information about the term's occurrence in the collection of documents;

(c) means for preparing a compressed document surrogate, for each document in the collection, said surrogate comprising a list of each term, for which there is an inverted term list, which occurs in the document, together with additional information about the occurrence of said term in said document;

(d) means for updating the set of inverted term lists, in response to a document in the collection which the set of inverted term lists summarizes being modified or deleted, by consulting the compressed document surrogate for said document to determine which terms for which there are inverted term lists in said document, and updating the inverted term lists corresponding to those terms to reflect: the modification or deletion in the document, and terms added to the document that were not previously in the document; and,

(e) means for specifying terms which are at least one of desired to be found in documents and which are desired not to be found, and for determining documents containing the desired terms and the undesired terms, and how often the desired terms and the undesired terms appear in the documents, by consulting the inverted term lists for the desired terms and the undesired terms, and preparing a list of documents ordered depending upon the occurrence of the desired terms or the undesired terms.

86. The device of claim 85, wherein a unique term identification number is assigned to each term, and a compressed document surrogate contains the term identification number of each term contained in the document.

87. The device of claim 86, wherein the information is stored in the compressed document surrogate in order of term identification number, and the term identification number of a term is given relative to the prior term identification number.

88. The device of claim 86, wherein a compressed document surrogate contains the number of times the term occurs in the document.

89. The device of claim 85, wherein the inverted term lists contain a document identification number for each document in which the term appears, and the number of times the term occurs in the document.

90. The device of claim 89, wherein documents are listed in an inverted term list in order of their term frequency scores.

91. The device of claim 85, wherein two inverted term lists are maintained for each term, a top inverted term list containing information about the documents in which the term occurs most frequently, and a remainder inverted term list containing information about the remaining documents in which the term occurs.

92. The device of claim 85, wherein a lookup table maintains, for each term, the term in a natural language, the address of each inverted term list for the term, the number of documents containing the term, and numbers reflecting the maximum amount the term can contribute to the score of a document on each of the term's inverted term lists, when processing a search query.

93. The device of claim 92, wherein the lookup table is a fixed array with information about terms stored in order of term identification numbers.

94. The device of claim 85, wherein an inverted term list for a term contains information about the location within a document of each occurrence of the term in question.

95. The device of claim 94, wherein the locations within the document of each occurrence of the term in question are given in relation to the prior occurrence of the term in the document.

96. The device of claim 85, wherein the documents are Web pages.

97. The device of claim 85, wherein the documents are Web sites.

98. A device for determining the score for a document under a search query which specifies terms that are to be present or absent, the device comprising:

(a) means for creating a compressed document surrogate for each document in the database, where the compressed document surrogate contains information about each term, from among the terms of interest of interest in the database, which occurs in the document, and the compressed document surrogate is created with inverted term lists that contain information about the terms of interest in the database, and where the information about each term included in the compressed document surrogate for a document includes at least one of: the term identification number of the term, the location in a lookup table of an entry for the term, the number of times the term occurs in the document, the location in the document of each occurrence of the term, the address of the inverted term list of the term, and the address of the location in the inverted term list for the term of the document;

(b) means for consulting the compressed document surrogate for the document whose score is to be determined;

(c) means for consulting at least one of an inverted term list and a lookup table, for each term contained in said compressed document surrogate, and calculating the contribution to the document score resulting from said term; and,

(d) means for determining the total document score by adding the contributions of each term in the compressed document surrogate.

99. The device of claim 98, where at (c) the inverted term list is not consulted.

100. The device of claim 98, wherein the documents are Web pages.

101. The device of claim 98, wherein the documents are Web sites.

102. A device for returning a list of a number of documents N in order of predicted utility, from among a collection of documents, as predicted by a search query containing terms to be present or absent, the device comprising:

(a) means for creating a compressed document surrogate for each document in the database, where the compressed document surrogate contains information about each term, from among the terms of interest of interest in the database, which occurs in the document, and where the compressed document surrogate is created with top and remainder inverted term lists that contain information about the terms of interest in the database, and where the information about each term included in the compressed document surrogate for a document includes at least one of: the term identification number of the term, the location in a lookup table of an entry for the term, the number of times the term occurs in the document, the location in the document of each occurrence of the term, the address of the inverted term list of the term which contains the document, and the address of the location in the inverted term list of the document;

(b) means for choosing, from among the terms in the search query which are to be found in documents, the term whose top inverted term list has not yet considered, which occurs in the fewest documents in the collection;

(c) means for consulting the top inverted term list for said term, and calculating the score for each document found in the top inverted term list;

(i) means for assigning the document the calculated score, in response to the document not having previously been found on an inverted term list;

(ii) means for increasing the document's previously-calculated score by the calculated score, in response to the document having previously been found on an inverted term list;

(d) means for calculating a maximum score, S_{Max}, achieved by a document, not already found on a top inverted term list, in response to it being found on all top inverted term lists, for terms to be found in documents, not yet consulted;

(e) means for calculating a maximum score, S_{Sub}, to be subtracted from a document score, as a result of said document being found to contain terms to be absent from a document;

(f) means for determining whether there are N or more documents already found, with scores such that if S_{Sub} were subtracted from their scores, the remainder would be greater than S_{Max};

(g) means for determining by use of the compressed document surrogate for each document a final score for the documents that have so far been found in any inverted term list of a desired term, and providing a list of the N documents with the highest scores, ranked in order of score, in response to there being N or more documents already found with scores such that if S_{Sub} were subtracted from their scores, the remainder would be greater than S_{Max};

(h) means for repeating (b) through (f) until either N or more such documents are found, or until no top inverted term list of a term desired to be found in the document has not been analyzed, in response to there not being N or more such documents;

(i) means for repeating (b) through (h) utilizing remainder inverted term lists instead of top inverted term lists, until either N or more such documents are found, or until no remainder inverted term lists of terms desired to be found in the document has not been analyzed, in response to there not being N or more such documents, and the top inverted term lists of all terms desired to be found in the document having been analyzed; and,

(j) means for determining by use of the compressed document surrogate for each document the final score for the documents found on the inverted term lists of the desired terms, and providing a list of the documents ranked in order of score.

103. The device of claim 102, wherein the documents are Web pages.

104. The device of claim 102, wherein the documents are Web sites.

105. The device of claim 102, wherein only terms desired to be found are contained in a search query, so that S_{Sub} is zero.

106. A device for choosing documents of interest from a collection of documents, the device comprising:

(a) means for determining an initial selection criterion;

(b) means for applying the initial selection criterion to each document in the collection, to generate a rank-ordered list of documents;

(c) means for evaluating a subset of the documents on the list to determine whether each document in the subset is relevant or irrelevant;

(d) means for modifying the selection criteria by at least one of: adjusting weights assigned to each element of the selection criteria in the prior iteration, removing elements of the selection criteria from the prior iteration, and adding additional elements to the selection criteria, based upon features of the relevant documents, by use of compressed document surrogates for the relevant documents, where said compressed document surrogates comprise information about the use of terms in the relevant documents;

(e) means for applying the modified selection criterion to each document in the collection, to generate a new rank-ordered list of documents; and,

(f) means for repeating (c), (d), and (e).

107. The device of claim 106, wherein when the modified selection criterion are applied to each document in the collection at (e), to generate a new rank-ordered list of documents; the compressed document surrogates for the documents are utilized to calculate the final document scores.

108. The device of claim 106, wherein the documents classified are Web pages.

109. The device of claim 106, wherein the documents classified are Web sites.

110. The device of claim 106, wherein the initial selection criteria are arbitrarily chosen.

111. The device of claim 106, wherein the documents classified are one of: electronic commerce Web pages and electronic commerce Web sites.

112. The device of claim 106, wherein means for modifying the selection criteria at (d) includes at least one of: means for adjusting a weight assigned to each element of the selection criteria in the prior iteration, means for removing elements of the selection criteria in the prior iteration, and means for adding additional elements to the criteria, based upon features of the irrelevant documents and the relevant documents, by use of compressed document surrogates for the relevant documents and the irrelevant documents , where said compressed document surrogates comprise information about the use of terms in the relevant documents and the irrelevant documents.

113. The device of claim 112, wherein means for modifying the selection criteria includes:

(a) means for giving each term found in the collection of documents a score based upon how often the term occurs in the relevant documents, compared to how often the term occurs in the collection of documents as a whole, and based upon how often the term occurs in the irrelevant documents , compared to how often the term occurs in the collection of documents as a whole;

(b) means for choosing terms with the highest positive weights thus determined to be the terms in the selection criteria; and,

(c) means for weighing the terms in the selection criteria according to the scores achieved in the above process, and the relative frequency of the terms in the collection.

114. The device of claim 113, wherein a score W_T given to a Term T at (a) is determined by a formula:

$W_T = \log(P_T(R)/P_T(\underline{R}))$, where

$P_T(R)$ = a probability that the term T occurs in a relevant document,

$= N_{TR}/(\Sigma_R N_{tR})$, where:

N_{TR} = a number of occurrences of the term T in a relevant document,

$\Sigma_R N_{tR}$ = a total number of occurrences of terms in relevant documents,

$P_t(\underline{R})$ = a probability that the term T occurs in an irrelevant document,

$= N_{T\underline{R}}/(\Sigma_R N_{t\underline{R}})$, where

$N_{T\underline{R}}$ = a number of occurrences of the term T in irrelevant documents,

$\Sigma_{\underline{R}} N_{t\underline{R}}$ = a total number of occurrences of terms in irrelevant documents.

115. The device of claim 114, wherein the terms chosen at (b) are the terms whose scores WT exceed an average score W_T by two or more standard deviations.

116. The device of claim 115, wherein weights S_T assigned to terms at (c) are determined by a formula:

$$S_T = W_T * IDF_T,$$

where: $IDF_T = \log((N+K_3)/N_T)/\log(N + K_4)$

where:

N is a number of documents in the subset,

N_T is a number of documents containing the term T in the subset,

K_3 and K_4 are constants.

117. The device of claim 116, wherein K_3 is 0.5, and K_4 is 1.0.

118. The device of claim 116, wherein in applying the modified selection criterion to each document in the collection, to generate a new rank-ordered list of documents, documents are ranked in order of their scores S_D,

where: $S_D = \Sigma(S_T * TF_{TD})$,

S_T has the value set forth above,

TF_{TD} = Robertson's term frequency for Term T in Document D

$= NT_D/(NT_D + K_1 + K_2*(L_D/L_0))$,

where: NTD is a number of times the term T occurs in document D,

L_D is a length of document D,

L_0 is an average length of a document in the subset of documents indexed, and

K_1 and K_2 are constants.

119. The device of claim 118, wherein K_1 is 0.5, and K_2 is 1.5.

120. A device for identifying documents in a collection as having a particular characteristic, the device comprising:

(a) means for choosing an initial list of documents from among the documents in the collection;

(b) means for evaluating a subset of the documents on the list to determine whether each document in the subset has the characteristic;

(c) means for modifying the selection criteria by at least one of: means for adjusting the weights assigned to each element of the selection criteria in the prior iteration, means for removing elements of the selection criteria in the prior iteration, and means for adding additional elements to the criteria, based upon features of the documents determined to have the characteristic, and based upon features of the documents determined not to have the characteristic, by use of compressed document surrogates for the documents, where said compressed document surrogates comprise information about the use of the terms in the documents found to have the characteristic or not to have the characteristic;

(d) means for applying the modified selection criterion to each document in the initial list of documents, to generate a new rank-ordered list of documents;

(e) means for repeating (b), (c), and (d);

(f) means for choosing a cutoff score to be applied; and,

(g) means for concluding that all documents in the collection with scores above the cutoff score have the characteristic.

121. The device of claim 120, wherein means for modifying the selection criteria at (c) includes:

(a) means for giving each term found in the subset of documents a score based upon how often the term occurs in documents determined to have the characteristic, compared to how often the term occurs in the subset of documents as a whole, and based upon how often the term occurs in documents determined not to have the characteristic, compared to how often the term occurs in the subset of documents as a whole;

(b) means for choosing terms with the highest positive weights thus determined to be the terms in the selection criteria; and,

(c) means for weighing the terms in the selection criteria according to the scores achieved in the above process, and their relative frequency in the subset.

122. The device of claim 121, wherein a score W_T given to a Term T at (a) is determined by a formula:

$W_T = \log(P_T(R)/P_T(\underline{R}))$, where

$P_T(R)$ = a probability that the term T occurs in a relevant document,

$= N_{TR}/(\Sigma_R N_{tR})$, where:

N_{TR} = a number of occurrences of the term T in a relevant document,

$\Sigma_R N_{tR}$ = a total number of occurrences of terms in relevant documents,

$P_t(\underline{R})$ = a probability that the term T occurs in an irrelevant document,

$= N_{T\underline{R}}/(\Sigma_R N_{t\underline{R}})$, where,

$N_{T\underline{R}}$ = a number of occurrences of the term T in irrelevant documents,

$\Sigma_{\underline{R}} N_{t\underline{R}}$ = a total number of occurrences of terms in irrelevant documents.

123. The device of claim 122, wherein the terms chosen at (b) are the terms whose scores W_T exceed an average score W_T by two or more standard deviations.

124. The device of claim 123, wherein weights S_T assigned to the terms at (c) are determined by a formula:

$S_T = W_T * IDF_T$,

where: $IDF_T = \log((N+K_3)/N_T)/\log(N + K_4)$

where:

N is a number of documents in the subset,

N_T is a number of documents containing the term T in the subset,

K_3 and K_4 are constants.

125. The device of claim 124, wherein K_3 is 0.5, and K_4 is 1.0.

126. The device of claim 124, wherein in applying the modified selection criterion to each document in the subset, to generate a new rank-ordered list of documents, documents are ranked in order of their scores S_D,

where: $S_D = \Sigma(S_T * TF_{TD})$,

S_T has the value set forth above,

TF_{TD} = Robertson's term frequency for Term T in Document D

$= NT_D/(NT_D + K_1 + K_2*(L_D/L_0))$,

where: NTD is a number of times the term T occurs in document D,

L_D is a length of document D,

L_0 is an average length of a document in the subset of documents indexed, and

K_1 and K_2 are constants.

127. The device of claim 126, wherein K_1 is 0.5, and K_2 is 1.5.

128. The device of claim 126, where the documents are Web pages.

129. The device of claim 126, where the documents are Web sites.

130. The device of claim 129, where the particular characteristic is being an electronic commerce site.